In: Wolfram Wilss, Klaus-Dirk Schmitz (Hrsg., 1986): **Maschinelle Übersetzung - Methoden und Werkzeuge**. Akten des 3. Internationalen Kolloquiums des Sonderforschungsbereichs 100 "Elektronische Sprachforschung", Saarbrücken 1.-3.9.1986, 33-48

MULTILINGUALE ANWENDUNGEN DER SPRACHDATENVERARBEITUNG IN REFERENZ-INFORMATIONSSYSTEMEN

Edith Kroupa, Harald H. Zimmermann

Abstract

Das Forschungsprojekt MARIS wird an der Fachrichtung Informationswissenschaft der Universität des Saarlandes mit Förderung des BMFT durchgeführt. Es hat den Aufbau und die exemplarische Erprobung einer computergestützten Übersetzung zum Ziel. Hierbei wird das am Sonderforschungsbereich elektronische Sprachforschung empirisch realisierte System SUSY (Saarbrücker Übersetzungssystem) mit herangezogen. Die Übersetzungen (Deutsch-Englisch) dienen der Überwindung von Sprachbarrieren in Fachinformationssystemen u.a. im Bauwesen, in der Normendokumentation und in den Sozialwissenschaften.

Es werden ein Überblick über die Entwicklungen von der maschinellen Indexierung (CTX) bis zum Modell des Saarbrücker Translations-Service (STS) gegeben und Ziele und Vorgehensweise vorgestellt.

Einleitung

Ein zentrales Problem, das sich dem Zugang zu in Informationsbanken (Datenbanken) gespeichertem Wissen entgegenstellt, sind die Sprach- und Kommunikationsbarrieren. Häufig beherrscht der Recherchierende die Sprache eines Dokuments bzw. die Indexierungssprache - sei es nun Fach- oder Fremdsprache - nicht oder nicht ausreichend.

Im fachsprachlichen Bereich muss man davon ausgehen, dass Informationen für einen Benutzer, der die Fachsprache überhaupt nicht beherrscht, auch relativ wertlos sind. Wenn dagegen ein Fachmann das spezielle Vokabular seines Fachs zumindest in gängigen Sprachen wie Englisch oder Französisch passiv beherrscht, liegen die (späteren) Probleme im Umgang mit fremdsprachigen Fachtexten weniger in der **Rezeption** der fremdsprachigen Texte als im **Such- und Wiedergewinnungsvorgang**.

Da die **rein** intellektuelle Übersetzung für große Mengen fremdsprachlicher Texte finanzielle und personelle Probleme aufwirft, bieten sich als Lösung die Übersetzung mit Computerunterstützung sowie eine Verbindung zwischen einsprachiger Indexierung und Übersetzungsäquivalenten des in den Fachtexten verwendeten Sprachmaterials an.

Im Folgenden wird die Sprachbarrierenproblematik am Beispiel von Referenz-Informationssystemen eingehender untersucht. Im ersten Kapitel wird eine kurze Übersicht über die Forschungsprojekte gegeben, in deren Zusammenhang die vorgestellten Systeme entwickelt wurden und werden.

1. Sprachdatenverarbeitung im Bereich Fachinformation - Forschungs- und Entwicklungsprojekte an der Fachrichtung Informationswissenschaft der Universität des Saarlandes

Aufbauend auf den Entwicklungsarbeiten des SFB 100 (vor allem des Teilprojekts A2) wurden zunächst an der Universität Regensburg und später an der Fachrichtung Informationswissenschaft der Universität des Saarlandes mehrere Forschungsprojekte im Bereich der Sprachdatenverarbeitung durchgeführt:

JUDO (1977/1979): Juristische Dokumentanalyse

Im Rahmen des Projekts JUDO wurde mit Finanzierung des BMFT an der Universität Regensburg auf der Grundlage des (damals noch in erster Entwicklung stehenden) Teilsystems zur deutschen Syntaxanalyse des SFB 100 ein Modell zur Dokumentation juristischer Texte entwickelt. Dabei wurden linguistisch motivierte Verfahren zur computergestützten Texterschließung konzipiert und erprobt. Die Erschließung erfolgte einschließlich der semantischen Disambiguierung und des Einsatzes differenzierter Thesaurusrelationen.

JUDO/DS (1980-1982): Juristische Dokumentanalyse im Bereich Datenschutz

Das prototypische Modell wurde in einem sich unmittelbar anschließenden Projekt, ebenfalls mit Finanzierung des BMFT, das noch an der Universität Regensburg begonnen, dann aber ab September 1980 an der Universität des Saarlandes weitergeführt wurde, an größeren Dokumentmengen in einem spezifischen Fachgebiet (insbesondere an Gesetzestexten, Berichten, Zeitungsmeldungen zum Datenschutz) getestet und fortentwickelt. Da sich bei JUDO herausstellte, dass die regelgeleitete Vereindeutigung innerhalb des Systems SUSY für die bearbeiteten Textsorten nicht ausreicht, wurden insbesondere in diesem Bereich neue Entwicklungen besonders zur Thesaurusintegration und zur Ausnutzung von Fachgebietsmarkierungen zur Disambiguierung durchgeführt. Am Ende des Projekts stand ein anwendbarer Prototyp zur Verfügung.

TRANSIT (1982/1985): Transfer informationslinguistischer Technologien

Im Projekt TRANSIT - auch hier wieder gefördert vom BMFT im Rahmen des Fachinformationsprogramms der Bundesregierung - wurde der in den Vorgängerprojekten entwickelte Prototyp auf andere Fachgebiete übertragen (Patentwesen, Werkstoffe, Sozialwissenschaften). Gleichzeitig wurde in jedem dieser Anwendungsgebiete eine spezielle Komponente weiterentwickelt. Bei den Patentdaten wurde die Praktikabilität bei großen Datenmengen getestet; gleichzeitig wurden aufgrund des Wortschatzes und der speziellen grammatikalischen Strukturen (lange, komplizierte Sätze) extreme Anforderungen an die Analysemoduln gestellt; bei den Sozialwissenschaften wurden die semantischen Komponenten getestet und z.T. weiterentwickelt, für das Fachgebiet Werkstoffe wurde ein Vergleich zwischen intellektuellen und maschinellen Indexierungsleistungen durchgeführt.

Ergebnisse von JUDO bis TRANSIT waren das Indexierungssystem CTX und eine erhebliche Stabilisierung der deutschen Analysekomponente von SUSY. Man kann dazu festhalten, dass in dieser Ausrichtung die SUSY-Teilkomponente 'Deutsche Sprachanalyse' für Indexierungszwecke ihre Einsatz- und vor allem ihre Leistungsfähigkeit deutlich gezeigt hat. Dies haben neutrale Be-

wertungen, z.B. im Vergleich einer unspezifischen Indexierung auf Wortformenebene zu SIEMENS/ PASSAT und SUSY/CTX, deutlich herausgestellt (Krause 1986).

SUSY-DJT (1983-1985): Deutsch-Japanische Titelübersetzung mit Englisch als Switching-Language

Das Projekt SUSY/DJT war ein deutsch-japanisches Kooperationsprojekt, an dem auf deutscher Seite neben der Fachrichtung Informationswissenschaft auch die Fachrichtung Übersetzen und Dolmetschen (Prof. Wilss) beteiligt war. Es handelte sich. um die labormäßige Übersetzung von Titeln japanischer Datenbanken aus dem Japanischen ins Deutsche und umgekehrt; dabei übersetzte die japanische Seite (Prof. Nagao in Kyoto) maschinell aus dem Japanischen ins Englische, mit SUSY wurden diese englischen Übersetzungen weiter ins Deutsche übersetzt und umgekehrt. Das Englische diente also zugleich als Switching Language. Die Schwierigkeiten bei der Übersetzung lagen zum einen in der Textsorte: Titel sind ohne Kontext bzgl. der korrekten Terminologie nicht leicht zu übersetzen; außerdem ist SUSY auf grammatikalisch korrekten Input ausgelegt, der bei den englischen Übersetzungen der japanischen Titel nicht immer gegeben war. Als Übersetzungsergebnis wurde allerdings lediglich eine Informativübersetzung angestrebt. Die Übersetzungsqualität wurde extern evaluiert (von Ammon, Wessely 1984; 1985) und erzielte (für alle) erstaunlich gute Ergebnisse. In diesem Projekt wurde auch eine zweisprachige Modelldatenbank mit zweisprachigen (englisch-deutschen) Deskriptoren aufgebaut.

MARIS (seit 1985): Multilinguale Anwendung von Referenz-Informationssystemen

Die Projekte SUSY/DJT und TRANSIT leiteten unmittelbar über zu dem Projekt MARIS, das seit Juni letzten Jahres unter Förderung des BMFT in Zusammenarbeit mit dem Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. (IAI) durchgeführt wird. Ziel von MARIS ist es, eine organisatorische und technische Infrastruktur für Übersetzungen im Fachinformationsbereich zu schaffen. Für die Übersetzungen (Deutsch-Englisch) wird wiederum SUSY eingesetzt. Bei der Textsorte handelt es sich um Titel aus unterschiedlichen Fachgebieten, z.Z. Technische Regeln, Bauwesen, Sozialwissenschaften. Ein Schwerpunkt ist dabei die Terminologieentwicklung. Erste Tests mit SUSY haben bereits gezeigt, dass unter der Voraussetzung der entsprechenden maschinenlesbaren Terminologie diese Textsorte sehr gut maschinell übersetzt werden kann. Ein Hauptvorteil der maschinellen bzw. computergestützten Übersetzung in diesem Bereich liegt in der terminologischen Konsistenz, die sich auf den späteren Rechercheverfahren sehr vorteilhaft auswirkt.

Wenn man die Verfahrensweisen und Ziele dieser außerhalb des SFB 100 'Elektronische Sprachforschung', aber in Kooperation mit Projektgruppen des SFB 100 durchgeführten Forschungsprojekte zusammenfasst, so lassen sich folgende Akzentuierungen festhalten:

Ziele des SFB 100 waren (u.a.), die prinzipiellen Möglichkeiten einer Sprachdatenverarbeitung theoretisch wie **empirisch** (an Modellen) zu erproben. Hierzu wurden Programme erstellt, Lexika kodiert und Grammatiken weitgehend systematisch implementiert und getestet. Der breitangelegte empirische Ansatz insbesondere beim Deutschen ermöglichte es, über die Modelle des SFB 100 hinaus Labor- und Prototypen (auch unter textbezogener Lexikonerweiterung) **extern** zu entwickeln.

Ziel dieser Anwendung war also weniger, die prinzipiellen Möglichkeiten zu testen, als zu prüfen, ob sich eine Anwendung (der entwickelten Verfahren) unter (letztlich praktischen) Anforderungen des Information Retrieval und der multilingualen Repräsentation 'lohnt', wobei auch Kosten-Nutzen-Relationen ökonomischer Art von Bedeutung waren. So 'lohnt' sich nach diesem Verständnis eine computergestützte Übersetzung dann, wenn sie kostengünstiger und/oder schneller als die sog. 'Humanübersetzung' erzielt werden kann. Fragen der Wartung, der Adaptierbarkeit, der Wörterbuchpflege stehen bei diesen Untersuchungen weit mehr im Vordergrund als die linguistischen Verfahren selbst.

2. Referenz-Informationssysteme

2.1 Referenzdatenbanken

Unter Referenz-Informationssystemen werden hier Informationssysteme verstanden, in denen Dokumentbeschreibungen gespeichert sind, die **Hinweise** geben auf die gesuchten Objekte. Diese Objekte sind in der Regel Zeitschriftenaufsätze, Monographien u.ä.; es kann sich jedoch auch um konkrete Gegenstände wie Gemälde, Ausgrabungsstücke etc. handeln. Insbesondere heißt dies, dass die Datenbanken nur Beschreibungen der gesuchten Objekte enthalten, mit deren Hilfe der Zugriff auf die Objekte ermöglicht wird, nicht jedoch die Objekte selbst. Nicht berücksichtigt werden hier also Fakten- und Volltextdatenbanken.

Die Beschreibung eines Objekts in einer bibliographischen Datenbank (Referenzdatenbank) setzt sich u.a. zusammen aus

- Titel
- (Abstract)
- Verfasser, (Herausgeber)
- (Serientitel)
- Erscheinungsjahr
- Deskriptoren, (Kategorien, Stichwörter)

Im gegebenen Zusammenhang werden ausschließlich die textuellen Teile einer Objektbeschreibung weiter untersucht. Diese sind

- Deskriptoren, begriffliche Klassifikationen
- Titel
- Abstract.

2.2 Indexierung und Retrieval

Der Input für die Datenbanken wird heute noch überwiegend intellektuell / manuell erstellt. Das bedeutet neben der intellektuellen Aufbereitung der bibliographischen Daten vor allem die intellektuelle Zuordnung von Suchbegriffen. Diese Suchbegriffe werden aus vorgegebenen Klassifikationen oder Thesauri bzw. aus kontrolliertem Vokabular ausgewählt. Daneben ist noch die so genannte Freitextrecherche möglich. Bei der Freitextrecherche wird mit Wörtern bzw. Wortfolgen aus den Textteilen der Dokumente gesucht.

Um verschiedene Wortformen zusammenzuführen, wird das Mittel der Trunkierung (Abschneiden von Wortteilen v.a. am Wortende) verwendet. Für das Englische bringt dies relativ gute Ergebnisse, bei stark flektierenden Sprachen wie dem Deutschen ergeben sich allerdings Probleme, z.B. bei den starken Verben und den umlautenden Pluralstämmen. Vor allem aber bei Wortzusammensetzungen führt die Trunkierung oft zu unerwünschten Ergebnissen. Im wesentlichen handelt es sich dabei um zwei Fehlerquellen:

- Es werden nicht alle Ausprägungen eines Lemmas erkannt; (Beispiel: Trunkierung von $MATRIX$ ($ = Zeichen für Begrenzung des Wortes) führt nicht zur unregelmäßigen Pluralform MATRIZEN. Dies bedeutet, dass die Trunkierung in diesen Fällen dazu führt, dass Belege nicht gefunden werden (= erniedrigter Recall).

- Es werden fälschlich Stämme zusammengeführt. Zum Beispiel erhält man bei beidseitiger Trunkierung von $SCHAFT$ neben VENTILSCHAFT auch STAATSANWALTSCHAFT. Dies führt zu Ballast und damit zu verminderter Präzision; die Auswirkung für den Recherchierenden besteht darin, dass er Zeit aufwenden muss, um die für ihn interessanten bzw. auf seine Fragestellung zutreffenden Ergebnisse auszusondern.

- Im besonderen Fall treffen beide Phänomene zu: LEXIKON müsste mit $LEXIK$ trunkiert werden, um auch die Pluralform LEXIKA einzubeziehen; bei $LEXIK$ werden aber auch Begriffe wie LEXIKOGRAPH, LEXIKOLOGIE etc. gefunden.

Der Recherchierende muss dann sowohl unzutreffende Dokumente aussondern, umgekehrt wird er eine Reihe zutreffender Texte nicht finden.

Um diesem Problem abzuhelfen, wurden z.B. für das Deutsche im Rahmen der hier beschriebenen Projekte entsprechende Lemmatisierungsverfahren entwickelt, die die Rückführung der Wortformen auf Grundformen und daneben eine Zerlegung von zusammengesetzten Wörtern in ihre sinnhaften Bestandteile ermöglichen.

Auswirkungen analog zur Trunkierung resultieren aus syntaktischen und semantischen Mehrdeutigkeiten eines Wortes, die mit den oben genannten Verfahren in der Regel jedoch nicht gelöst werden. In Kapitel 5 werden Ansätze beschrieben, die auch für dieses Problemfeld Lösungsmöglichkeiten anbieten.

2.3 Mehrsprachigkeit

Bei wissenschaftlichen Veröffentlichungen nimmt der Anteil an englischsprachigen Publikationen ständig zu. Es zeigt sich, dass sich Englisch als internationale Wissenschaftssprache weitgehend durchgesetzt hat. Für den deutschen Fachinformationsmarkt zeichnen sich dadurch nun zwei gegenläufige Entwicklungstendenzen ab. Einerseits ergibt sich für Anbieter von deutschsprachiger Fachinformation die Notwendigkeit, Übersetzungen ins Englische vorzunehmen, um so die internationale Akzeptanz der eigenen Informationen zu erhöhen und damit das Marktsegment über den deutschsprachigen Markt hinaus auszuweiten. Umgekehrt stellt sich für die deutschsprachige Nutzung das Problem, dass der Informationssuchende (Experte) in der Regel zwar über ausreichend gute **passive** englische Sprachkenntnisse verfügt, um englischsprachige Texte zu verstehen, in englischsprachigen Datenbeständen jedoch nur dann die gewünschten In-

formationen findet, wenn er ausreichende **aktive** Sprachkenntnisse besitzt (von der Einschaltung eines Informationsvermittlers sei hier abgesehen).

Daraus resultieren heute 'doppelt' geführte Datenbanken, wenn ein Anbieter auf den deutschen Markt und daneben auf internationale Nutzung ausgerichtet ist (Wissmann 1986). Die 'deutsche' Datenbank stellt in der Regel eine Teilmenge der 'internationalen' Datenbank dar. Von der Marktlage her ist also durchaus ein Bedürfnis nach (kostengünstigen) Übersetzungen im Fachinformationsbereich festzustellen.

3. Multilinguale Objektbeschreibung

Der am ehesten zu realisierende multilinguale Zugang zu Datenbanken kann über eine mehrsprachige (zunächst auch intellektuelle) Indexierung erfolgen. Bei der Datenbank SDIM-2 des Fachinformationszentrums Werkstoffe verfolgt man z.B. diesen Weg. In diesen Datenbanken werden heute schon Deskriptoren in drei Sprachen (Deutsch, Englisch, Französisch) und Titel in mindestens zwei Sprachen (Deutsch, Englisch) angeboten (SDIM-2 1985).

3.1 Deskriptoren in mehreren Sprachen

Bezieht man mehrsprachige Indexierung sowohl auf kontrolliertes Vokabular als auch auf Einträge in den Basic Indices, liegt das Problem bei der Automatisierung der Zuordnung auch hier in der Auflösung von Mehrdeutigkeiten. Ist z.B. im Ausgangsvokabular der Begriff Spannung mit voltage (im Elektrotechnikbereich) und mit stress (in der Mechanik: Werkstoffbereich) übersetzt, so würde der menschliche Indexierer aufgrund seiner Fachkenntnisse den zutreffenden Ausdruck wählen; vom Programm aber kann die zutreffende Bedeutung ohne weitere Markierungen nicht automatisch erkannt werden; der Algorithmus 'merkt' nicht einmal in jedem Fall, ob es sich überhaupt um einen mehrdeutigen Begriff handelt. Deshalb müssten bei der ausgangssprachigen Indexierung potentiell mehrdeutige Begriffe bereits markiert werden.

Damit ist dann zwar eine eindeutige Zuordnung der zielsprachigen Begriffe zu realisieren, für den quellsprachig recherchierenden Benutzer bestehen zwei Möglichkeiten: Werden die Unterscheidungsmerkmale getilgt, so ändert sich für den Benutzer an der Oberfläche nichts, allerdings erzeugt dann die Bedeutungsvielfalt beim Rechercheergebnis Ballast (unzutreffende Dokumente). Wird das Unterscheidungsmerkmal mitgeführt, führt dies beim Benutzer zu einer aufwendigeren Handhabung des Rechercheapparats (allerdings sind auch hier technische Abhilfen denkbar).

3.2 Übersetzung von Titeln und Abstracts

Dem Nutzer wird durch die Übersetzung von Deskriptoren (Suchwortbegriffen) nur der erste Zugang zur Information erleichtert; ergeben sich zu den Stichwörtern Treffer, so hat er bei fremdsprachigen Texten immer noch das Problem, die Informationen zu verstehen. Dies hat - wie erwähnt - insbesondere Auswirkungen auf die Akzeptanz von nicht-englischsprachigen Informationen: denn umgekehrt hat sich zur Zeit wohl Englisch als Wissenschaftssprache durchgesetzt, so dass englischsprachige Informationen zumindest passiv rezipiert werden können.

Nachdem der Nutzer ein fremdsprachiges Dokument gefunden hat, wird ihm die Entscheidung über die Relevanz leichter fallen, wenn ihm Titel oder besser noch Titel und Abstract in einer Sprache angeboten werden, die er versteht.

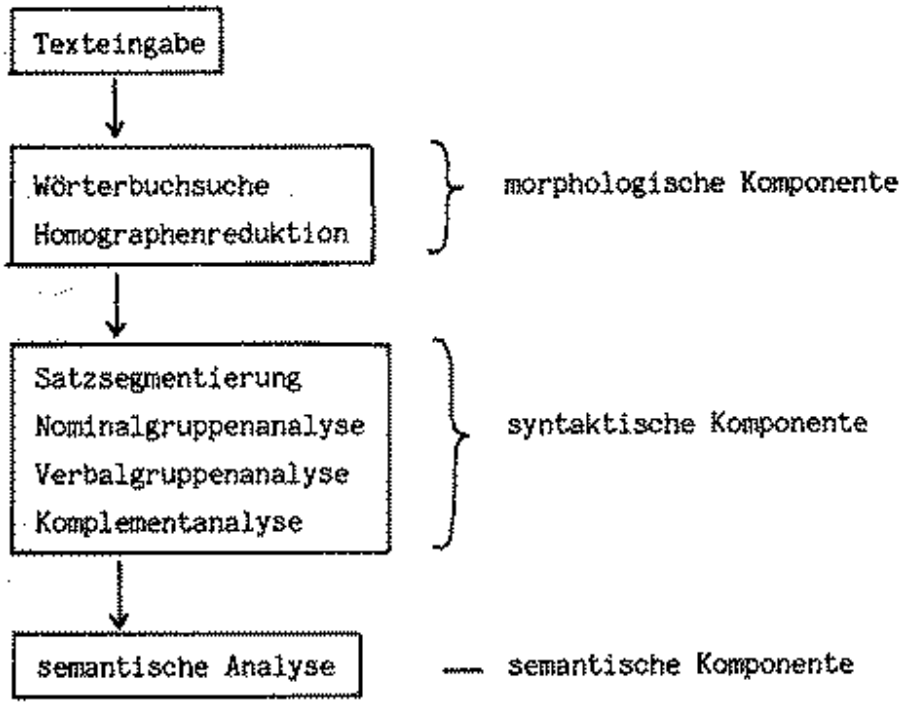
Inzwischen sind maschinelle Übersetzungshilfen entwickelt worden, die in einem gewissen Grade eine rationellere Durchführung der Übersetzungsaufgaben ermöglichen (Einsparungen von 30% und mehr an Übersetzungszeit sind durchaus möglich). Ihr Einsatz lohnt sich jedoch erst ab einem gewissen Mindestvolumen an Übersetzungen. Diese Hilfsmittel (Systeme wie SYSTRAN oder LOGOS) unterstützen die fremdsprachige Textgenerierung (Übersetzung, Postedition) und führen im terminologischen Bereich zu größerer Konsistenz.

4. Das maschinelle Übersetzungssystem SUSY und das computergestützte Texterschließungssystem CTX

Das Saarbrücker Übersetzungssystem SUSY und das computergestützte Texterschließungssystem CTX sind anderweitig ausführlich beschrieben (Vgl. Zimmermann, Kroupa, Keil 1983). Deshalb sei hier zum besseren Verständnis nur ein kurzer Abriss gegeben.

SUSY ist ein modulares, multilinguales Übersetzungssystem, dreigeteilt in Analyse, Transfer und Synthese.

Die ausgangssprachliche **Analyse** besteht aus den Moduln:



An die semantische Komponente schließt der lexikalische und strukturelle Transfer an.

Die zweisprachige Generierung besteht aus semantischer, syntaktischer und morphologischer Synthese.

CTX erzeugt unter Verwendung von SUSY-Teilen in Verbindung mit zusätzlichen Bausteinen aus Wortformen Grundformen (in unterschiedlichen Analysetiefen) und bildet syntaktische Relationen auf zweistellige syntaktisch relationierte Suchbegriffe ab. CTX nutzt also die Analysestufe von SUSY aus. Je nach Analysetiefe sind entsprechende Indexierungsergebnisse möglich.

SUSY / ERGÄNZUNGEN	CTX
morphologische Komponente	Grundformen, Kompositumzerlegung
syntaktische Komponente	Mehrwortbegriffe
semantische Komponente	semantische Vereindeutigung

Zusätzlich ist normalerweise in CTX ein Textpräprozessor vorgeschaltet, der eine logische Einteilung der Texte in kleinere Einheiten ermöglicht und die Informationen für die spätere Deskriptor-Text-Zuordnung bereitstellt.

Weiterhin wurden in Modellform zusätzliche Verfahren zur semantischen Vereindeutigung entwickelt, die Kontext und Fachgebietsmarkierungen auswerten.

Die folgende Gegenüberstellung zeigt noch einmal im Detail, wie die Ergebnisse der einzelnen Analysemodule des Übersetzungssystems zur Lösung einsprachiger Dokumentationsaufgaben genutzt werden können.

Maschinelle Übersetzung		Dokumentation
		Dokumentstruktur
Wort-/Textsegmentierung	{	Rechtschreibfehler/ Wortzerlegungen
Morphologische Analyse	{	Grundformen Stoppwörter Trunkierung
Satzsegmentierung	—	Abstandsfunktionen
syntaktische Analyse	—	nominale Mehrwortgruppen syntaktische Phrasen
semantische Analyse	—	Auflösung von Mehrdeutigkeiten

Das Verfahren lässt sich auf ein- und mehrwortiger Begriffsebene noch erweitern: Das Transfermodul liefert die zielsprachigen Entsprechungen zu den ausgangssprachigen Begriffen. Mit dem Ergebnis des gesamten Übersetzungsprozesses kann schließlich eine komplette zweisprachige Datenbank aufgebaut werden.

Der Zugang zu fremdsprachigen Informationen kann damit auf zwei Arten ermöglicht werden: Einmal durch die formal-inhaltliche Erschließung von Texten und die Relationierung der gewonnenen Suchbegriffe und fremdsprachigen Äquivalenten; zum Zweiten (z.Z. erheblich weniger effizient) durch die maschinelle Übersetzung des ausgangssprachigen Textes und die formal-inhaltliche Erschließung des zielsprachigen Textes.

5. Integriertes Terminologie- und Übersetzungskonzept

Im Fachinformationsbereich ergeben sich jedoch gerade bei der Übersetzung von Deskriptoren und Titeln eine Reihe spezifischer Probleme.

Informationsbanken zu bestimmten Themenbereichen decken in sich wieder ein größeres Spektrum an Fachgebieten ab (extreme Beispiele: Technische Regeln, Patente). Dies hat entsprechende Auswirkungen auf den Umfang des zu bearbeitenden Wortschatzes. Für das Bauwesen (Beispiel: die Datenbank des Informationszentrums Raum und Bau (IRR) in Stuttgart) hat sich z.B. ergeben, dass erst ab ca. 100.000 bearbeiteten Titeln eine gewisse Sättigung zu erwarten ist. Für den Einsatz maschineller Übersetzungsverfahren bedeutet dies, dass für jedes neu zu bearbeitende 'Fachgebiet' umfangreiche terminologische Vorarbeiten erforderlich sind.

Mehrdeutigkeiten/Fachgebietsmarkierung

Mehrdeutigkeiten können nur im Kontext aufgelöst werden; bei der Übersetzung isolierter Termini und auch noch bei Titeln reicht der Kontext oft nicht aus, um Mehrdeutigkeiten erkennen, geschweige denn auflösen zu können. Handelt es sich um bereits intellektuell klassifizierte Dokumenteinheiten, können die Fachgebietszuweisungen ausgewertet werden. Sie bieten sich auch zur automatischen Klassifizierung der in den Texten auftretenden Termini an. Somit erhält man ein lernendes Verfahren zur automatischen Disambiguierung, das allerdings auf intellektuelle Vorarbeit angewiesen ist.

Terminologiegenerierung

Der Neuheitswert der Information äußert sich auch in den Benennungen der untersuchten Gegenstände; für den Terminologen bzw. Übersetzer bedeutet dies oft, dass die fremdsprachigen Entsprechungen mühsam recherchiert werden müssen.

Eine wesentliche Aufgabe im Rahmen der praxisorientierten Entwicklung eines maschinellen Übersetzungssystems ist somit der Ausbau der Wörterbücher und Fachterminologien sowie deren Anpassung an die vom Übersetzungssystem benötigten Kodierungen. Ein Beispiel für entsprechende Ergebnisse ist im Saarbrücker Übersetzungssystem das morphosyntaktische Wörterbuch zum Deutschen. Es umfasst bereits aufgrund der Basis-Entwicklungen im SFB 100 und den textorientierten Ergänzungen durch die Indexierungsprojekte ca. 150.000 Einträge und deckt den größten Teil des deutschen Funktionswortschatzes und des deutschen Standardwortschatzes ab. Für die praktischen Übersetzungen muss für die zu bearbeitenden Fachgebiete der entsprechende Aufbau der Transfer- und Synthesewörterbücher folgen.

Es werden verschiedene Strategien verfolgt, die den Ausbau der bestehenden Lexika (v.a. Deutsch/Englisch) kostengünstig gestalten bzw. als Beiprodukt ermöglichen. Sie werden in den

Entwicklungen des sog. Saarbrücker Translations-Service (STS) im Rahmen des Projekts MARIS erprobt.

5.1 Übernahme vorhandener maschinenlesbarer Terminologie

Bei exemplarischen Untersuchungen (Peters 1985; Button 1986) hat sich gezeigt, dass aufgrund der zu bearbeitenden Textsorte (Titel von Neuerscheinungen in den jeweiligen Fachgebieten) mit relativ vielen unterschiedlichen Begriffen zu rechnen ist, die bislang weder in Fachwörterbüchern noch in Terminologiebanken zu finden sind. Es gibt z.B. bis dato keine Terminologiedatenbank, die speziell auf das Bauwesen ausgerichtet ist. Zudem ist dieser Themenbereich sehr weit zu fassen. Auch bei den technischen Regeln und Vorschriften lässt sich im eigentlichen nicht von **einem** Fachgebiet sprechen. Die Übernahme von Terminologie ist nicht notwendig kostengünstiger, ganz abgesehen von den weitgehend ungeklärten Rechtsfragen (Copyright). Die Übernahme vorhandener maschinenlesbarer Terminologie im Rahmen von MARIS beschränkt sich deshalb bislang i.W. auf die Übernahme der Bestände, die bei den einzelnen Projektpartnern, für die Übersetzungen ausgeführt werden, bereits vorliegen. Dabei handelt es sich bislang insbesondere um:

- Deutsches Informationszentrum RAUM und BAU:
  FINDEX Facettenorientiertes Indexierungssystem (ca. 6.800 Begriffe);

- Informationszentrum für technische Regeln:
  deutsch/englische Registerbegriffe (ca. 11.000 Begriffe).

5.2 Auswertung von übersetztem Textmaterial

Zweisprachiges Textmaterial ist für die computergestützte Terminologiegenerierung interessant, wenn es bereits maschinenlesbar vorliegt. Im vorliegenden Falle handelt es sich im Wesentlichen um zweisprachig vorliegende Datenbestände in den Datenbanken, für die mit STS Serviceleistungen erbracht werden (für anderes maschinenlesbares Textmaterial gelten bzgl. der Verfügbarkeit und des Copyright analoge Aussagen wie für den Terminologiebereich). Aus personellen und terminlichen Gründen war es bislang nicht möglich, mehrsprachiges maschinenlesbares Material, das **nicht** im Rahmen von STS übersetzt wird, aufzuarbeiten, obwohl die entwickelte Verfahrensweise auch für Fremdmaterial einsetzbar wäre. Somit erfolgt die Terminologieerstellung vorerst parallel zu den Übersetzungsarbeiten. Die übersetzten Texte werden mittels der computergestützten Texterschließung der deutschen Quelltexte durch CTX (Kroupa 1984) ausgewertet und anschließend in einer Terminologiedatenbank (mit dem Datenbanksystem GOLEM) abgelegt. Über Thesaurusrecherchen werden zu jedem Begriff die zweisprachigen Kontexte ermittelt, die dann von den Übersetzern für die Zuordnung von Übersetzungsäquivalenten ausgewertet werden. Als Seiteneffekt des gewählten Verfahrens entsteht auf der Basis der übersetzten Daten eine Terminologiedatenbank, die von den Übersetzern bei Problemfällen als Kontext herangezogen werden kann.

In Abhängigkeit vom Ausgangsmaterial werden bei der Terminologieerstellung drei Stufen unterschieden. Ziel ist immer eine hoch qualifizierte Terminologieerstellung.

Bei TERM 1 liegen zu den quellsprachigen Ausgangsbegriffen bereits die Übersetzungen vor; diese müssen lediglich noch auf die Systemformate umgesetzt werden.

Bei TERM 2 fehlen zwar die Übersetzungsäquivalente zu den Ausgangsbegriffen, es liegt zu jedem Ausgangsbegriff ein quellsprachiger und ein diesem zugeordneter zielsprachiger Kontext vor.

Bei TERM 3 kann der Kontext zu den Ausgangsbegriffen fehlen oder nur teilweise vorhanden sein, außerdem fehlt i.d.R. die Zuordnung.

5.3 Übersetzungskonzepte

Auch für Referenzdatenbanken sind i.W. zwei Arten von Übersetzungen denkbar. Orientiert man sich nur an dem Informationsbedürfnis des Benutzers, so reicht eine sog. Informativübersetzung (= Rohübersetzung) aus. Solche Übersetzungen werden z.T. auch von Übersetzungsabteilungen von Fachinformationseinrichtungen seit längerer Zeit intellektuell angefertigt. Der Kunde formuliert sein Informationsbedürfnis und erhält eine Zusammenfassung oder Übersetzung von Teilen, die nicht bis ins Detail ausformuliert zu sein braucht.

Aus Marketinggründen sollen sich aber übersetzte Dokumentationseinheiten von Originaleinträgen weder stilistisch noch terminologisch unterscheiden; dies erfordert eine hochqualitative Übersetzung, die dennoch aus Kostengründen preiswert gehalten werden soll.

Dazu bieten sich maschinelle Verfahren an. Es hat sich (zumindest bei Tests von SUSY) gezeigt, dass bei entsprechenden Terminologievorarbeiten gerade für Titel von Fachveröffentlichungen durchaus akzeptable maschinelle Übersetzungen, die wenig Nachbearbeitung erfordern, angefertigt werden können.

Liegt die entsprechende Terminologie nicht maschinenlesbar vor, so bietet es sich aus Effizienzgründen an, Texte zuerst herkömmlich zu übersetzen und anschließend zur Terminologieauswertung zu verwenden. Bei ausreichender terminologischer Sättigung wird das Verfahren umgekehrt: Wörterbücher werden nur noch bei lexikalischen Lücken, die bei der Übersetzung auftreten, ergänzt.

'Abfallprodukt' des automatischen Übersetzungsprozesses ist dabei (bei der Kopplung von SUSY und CTX zumindest), eine automatische Indexierung und die Zuordnung von fremdsprachigen Übersetzungsäquivalenten zu den Deskriptoren. Bei Bedarf lässt sich aus den so verarbeiteten Textteilen auch eine Terminologiedatenbank für Übersetzer aufbauen.

Die ersten Erfahrungen im Rahmen des Projekts MARIS zeigen, dass die Strategie bislang problemlos aufgeht: Inzwischen sind allein im Bereich des Bauwesens (Datenbank ICONDA) rund 80.000 Titel übersetzt und in ihrer Terminologie inventarisiert worden. Die entsprechend 'angereicherte' Datenbank wird soeben in den USA durch den Anbieter präsentiert.

Insgesamt ist festzuhalten, dass aus der Sicht der Informationsverarbeitung mit den Saarbrücker Arbeiten auf der Grundlage der empirischen Forschungen des SFB 100 zur Sprachdatenverarbeitung deutliche Fortschritte gemacht werden konnten. Mit dem Dank an den SFB 100 für die gute Zusammenarbeit wird zugleich die Hoffnung verbunden, dass weitere Entwicklungen - wie

sie beispielsweise in Saarbrücken mit EUROTRA begonnen wurden - zu ähnlichen Möglichkeiten führen werden.

LITERATUR

Button, D. (1986): Synoptischer Wörterbuchvergleich. Saarbrücken.

Krause, J. (Hg.) (1986): Inhaltserschließung von Massendaten. Zur Wirksamkeit informationslinguistischer Verfahren am Beispiel des Deutschen Patentinformationssystems. Erscheint (Sprache und Computer 8). Hildesheim, Zürich, New York.

Kroupa, E. (1983): Die Demonstrationsdatenbank SUSY-DJT - Beschreibung und Benutzeranleitung. Saarbrücken.

Peters, J.-P. (1985): Übersicht über multilinguale Terminologie. (masch.) Saarbrücken.

SDIM-2 (1985): Hg. vom INKA Online Service (Bluesheet). Karlsruhe.

von Ammon, R., R. Wessoly (1984): Das Evaluationskonzept des automatischen Übersetzungsprojektes SUSY-DJT (Deutsch-Japanische Titelübersetzung), Teil 1. In: Multilingua Vol. 3-4/1984. 189ff.

von Ammon, R., R. Wessoly (1985): Das Evaluationskonzept des automatischen Übersetzungsprojektes SUSY-DJT (Deutsch-Japanische Titelübersetzung), Teil 2. In: Multilingua Vol. 4-1/1985. 27ff.

Wißmann, W. (1980): ICONDA - Die erste internationale Datenbank für das Bauen und Planen. In: Nachrichten für Dokumentation Vol. 3/1980. 168ff.

Zimmermann, H.H., E. Kroupa, G. Keil (1983): CTX. Ein Verfahren zu computergestützten Texterschließung. (BMFT-FB-ID83-006). Karlsruhe.